CHAPTER 2

Introduction to Probability

1. Definitions

In weather forecasts one often hears a sentence such as "the probability of rain tomorrow is 50 percent." What does this mean? Something like: "If we look at all possible tomorrows, in half of them there will be rain", or "if we make the experiment of observing tomorrow, there is a quantifiable chance of having rain tomorrow, and somehow or other this chance was quantified as being 1/2". To make sense of this, we formalize the notions of experimental outcome, event, and probability.

Suppose that you make an experiment and imagine all possible outcomes.

DEFINITION. A sample space Ω is the space of all possible outcomes of an experiment.

For example, if the experiment is "waiting until tomorrow, and then observing the weather", Ω is the set of all possible weathers of tomorrow. There can be many weathers, some differing only in details we cannot observe, and with many features we cannot describe precisely.

Suppose you set up a thermometer in downtown Berkeley and decide you will measure the temperature tomorrow at noon. The set of possible weathers for which the temperature is between 65 and 70 degrees is an "event", an outcome which is specified precisely and about which we can think mathematically. An event is subset of Ω, a set of outcomes, a subset of all possible outcomes Ω, all of which share a well-defined property which can be measured.

DEFINITION. An event is a subset of Ω.

The set of events we are able to consider is denoted by $\mathcal{B}$; it is a set of subsets of Ω. We require that $\mathcal{B}$ (the collection of events) be a σ-algebra, i.e., $\mathcal{B}$ satisfy the following axioms:

(1) $\emptyset \in \mathcal{B}$ and $\Omega \in \mathcal{B}$ ($\emptyset$ is the empty set).
(2) If $B \in \mathcal{B}$ then $CB \in \mathcal{B}$ (CB is the complement of B in Ω).

(3) If $\mathcal{A} = \{A_1, A_2, \ldots, A_n, \ldots\}$ is a finite or countable collection in $\mathcal{B}$, then any union of the elements of $\mathcal{A}$ is in $\mathcal{B}$.

It follows from these axioms that any intersection of a countable number of elements of $\mathcal{B}$ also belongs to $\mathcal{B}$.

Consider the tosses of a die. In this case, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

(1) If we are only interested in whether something happened or not, we may consider a set of events

$$\mathcal{B} = \{\{1, 2, 3, 4, 5, 6\}, \emptyset\}.$$

The event $\{1, 2, 3, 4, 5, 6\}$ means "something happened" while the event $\emptyset$ means "nothing happened".

(2) If we are interested in whether the outcome is odd or even then we may choose

$$\mathcal{B} = \{\{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}, \emptyset\}.$$

(3) If we are interested in which particular number appears then $\mathcal{B}$ is the set of all subsets of Ω. In this case we can say that $\mathcal{B}$ is generated by $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.

Observe that $\mathcal{B}$ in case (1) is the smallest σ-algebra on the sample space (in the sense of having fewest elements) while $\mathcal{B}$ in case (3) is the largest.

DEFINITION. A probability measure $P(A)$ is a function $P : \mathcal{B} \to \mathbb{R}$ defined on the sets $A \in \mathcal{B}$ such that

(1) $P(\Omega) = 1$.
(2) $0 \leq P \leq 1$.
(3) If $\{A_1, A_2, \ldots, A_n, \ldots\}$ is a finite or countable collection of events such that $A_i \in \mathcal{B}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ (the probability of the simultaneous occurrence of incompatible events is the sum of the probabilities of the individual events).

DEFINITION. The triple $(\Omega, \mathcal{B}, P)$ is called a probability space.

In brief, the σ-algebra $\mathcal{B}$ defines the objects to which we assign probabilities and P assigns probabilities to the elements of $\mathcal{B}$.

DEFINITION. A random variable $\eta : \Omega \to \mathbb{R}$ is a $\mathcal{B}$-measurable function defined on Ω where "$\mathcal{B}$-measurable" means that the subset of elements ω in Ω for which $\eta(\omega) \leq x$ is an element of $\mathcal{B}$. In other words, it is possible to assign a probability to the occurrence of the inequality $\eta \leq x$ for every x.

Loosely speaking, a random variable is a real variable whose numerical values are determined by experiment, with the proviso that it is possible to assign probabilities to the occurrence of the various values.

Given a probability measure $P(A)$, the probability distribution function of a random variable η is given by

$$F_\eta(x) = P\left(\{\omega \in \Omega \mid \eta(\omega) \leq x\}\right) = P(\eta \leq x).$$

The existence of such a function is guaranteed by the definition of a random variable.

Now consider several examples.

Example. Let $\mathcal{B} = \{A_1, A_2, A_1 \cup A_2, \emptyset\}$. Let $P(A_1) = P(A_2) = 1/2$. Define a random variable

$$\eta(\omega) = \begin{cases} -1, & \omega \in A_1 \\ +1, & \omega \in A_2 \end{cases}.$$

Then

$$F_\eta(x) = \begin{cases} 0, & x < -1 \\ 1/2, & -1 \leq x < 1 \\ 1, & x \geq 1 \end{cases}.$$

Example. Suppose that Ω is the real line and the range of a random variable η also is the real line, e.g., $\eta(\omega) = \omega$. In this case one should be sure that the σ-algebra $\mathcal{B}$ is large enough to include all of the sets of the form $\{\omega \in \Omega \mid \eta(\omega) \leq x\}$. The minimal σ-algebra satisfying this condition is the σ-algebra of the Borel sets formed by taking all the possible unions and complements of all of the half-open intervals in $\mathbb{R}$ of the form $(a, b]$.

Example. Suppose that we are tossing a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\eta(\omega) = \omega$. Take $\mathcal{B}$ to be the set of all subsets of Ω. The probability distribution function of η is the shown in Figure 1.

Now for any $a, b \in \mathbb{R}$, with $a < b$ one can find c such that $a < c \leq b$ and

$$P(c) = F_\eta(b) - F_\eta(a).$$

Suppose that $F'_\eta(x)$ exists. Then $f_\eta(x) = F'_\eta(x)$ is the probabilty density of η. Since $F_\eta(x)$ is non-decreasing $f_\eta(x) \geq 0$. Obviously,

$$\int_{-\infty}^{\infty} f_\eta(x)dx = F_\eta(\infty) - F_\eta(-\infty) = 1.$$

If $F'_\eta(x)$ exists and is continuous then

$$P(x < c \leq x + dx) = F_\eta(x + dx) - F_\eta(x) = f_\eta(x)dx.$$

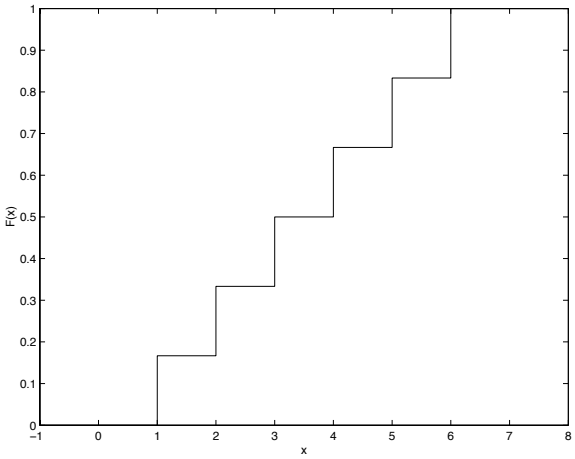

FIGURE 1. Probability distribution for a fair six sided die.

The following probability density functions (pdf's) are often encountered:

(1) Equidistribution density

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

(2) Gaussian density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{2.1}$$

where m and σ are constants.

(3) Exponential density

$$f(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & x \le 0 \end{cases}.$$

2. Expected Values

DEFINITION. Let $(\Omega, \mathcal{B}, P)$ be a probability space and η be a random variable. Then the expected value, or mean, of the random variable η is defined as the integral of η over Ω with respect to the measure P:

$$E[\eta] = \int_\Omega \eta(\omega) dP.$$

In the case where Ω is a discrete set this integral is just the sum of the products of the values of η with the probabilities that η assume these values.

This definition can be rewritten in another way involving the Stieltjes integral. Let F be a non-decreasing and bounded function. Define the Stieltjes integral of a function $g(x)$ on an interval $[a, b]$ as follows. Let $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$, $\Delta_i = x_{i+1} - x_i$, and $x_i^* \in (x_i, x_{i+1})$. Then

$$\int_a^b g(x)dF(x) = \lim_{\Delta_i \to 0} \sum_{i=1}^{n} g(x_i^*)(F(x_{i+1} - F(x_i))$$

(where we have written F instead of F_η for short). The expected value of η is

$$E[\eta] = \int_{-\infty}^{\infty} x dF(x).$$

If the derivative $F' = f(x)$ exists then

$$E[\eta] = \int_{-\infty}^{\infty} x f(x) dx.$$

If η is a random variable then so is $a\eta$ where a is a constant. If η is a random variable and $g(x)$ is a continuous function defined on the range of η, then $g(\eta)$ is also a random variable, and

$$E[g(\eta)] = \int_{-\infty}^{\infty} g(x) dF(x).$$

The special cases:

$$E[\eta^n] = \int_{-\infty}^{\infty} x^n dF(x)$$

and

$$E[(\eta - E[\eta])^n] = \int_{-\infty}^{\infty} (x - E[\eta])^n dF(x)$$

are called the n^{th} moment and the n^{th} centered moment of η respectively. (Of course, these integrals may fail to converge for some random variables.) The 2^{nd} centered moment is the variance of η.

DEFINITION. The variance $\mathrm{Var}(\eta)$ of the random variable η is

$$\mathrm{Var}(\eta) = E[(\eta - E[\eta])^2]$$

and the standard deviation of η is

$$\sigma = \sqrt{\mathrm{Var}(\eta)}.$$

EXAMPLE. The Gaussian pdf (2.1) has $E[\eta] = m$ and $\mathrm{Var}(\eta) = \sigma^2$.

DEFINITION. Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Two random variables η_1 and η_2 are independent if the events $\{\omega \in \Omega \mid \eta_1(\omega) \leq x\}$ and $\{\omega \in \Omega \mid \eta_2(\omega) \leq y\}$ are independent for all x and y.

DEFINITION. If η_1 and η_2 are random variables then the joint distribution function of η_1 and η_2 is defined by

$$F(x, y) = P(\{\omega \in \Omega \mid \eta_1(\omega) \leq x, \eta_2(\omega) \leq y\}) = P(\eta_1 \leq x, \eta_2 \leq y).$$

If the second mixed derivative $\partial^2 F(x, y)/\partial x \partial y$ exists then it is called the joint density of η_1 and η_2 and denoted by $f_{\eta_1 \eta_2}$. In this case

$$F_{\eta_1 \eta_2}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\eta_1 \eta_2}(s, t)\, ds\, dt.$$

Clearly if η_1 and η_2 are independent then

$$F_{\eta_1 \eta_2}(x, y) = F_{\eta_1}(x) F_{\eta_2}(y)$$

and

$$f_{\eta_1 \eta_2}(x, y) = f_{\eta_1}(x) f_{\eta_2}(y).$$

We can view two random variables η_1 and η_2 as a single vector valued random variable $\eta = (\eta_1, \eta_2) = \eta(\omega)$ for $\omega \in \Omega$. Then η is measurable if the event $\eta \in S$ with $S \in \mathbb{R}^2$ is measurable for a suitable family of S's, i.e., the event $Z = \{\omega \in \Omega : \eta(\omega) \in S\} \in \mathcal{B}$, where $\mathcal{B}$ is a σ-algebra on Ω. Suppose that the joint probability function of the two random variables exists and is denoted by $F_{\eta_1 \eta_2}(x, y) = P(\eta_1 \leq x, \eta_2 \leq y)$. Note that $F_{\eta_1 \eta_2}(x, y) = F_{\eta_2 \eta_1}(y, x)$ and $F_{\eta_1 \eta_2}(\infty, y) = F_{\eta_2}(y)$. If the density of the joint probability distribution exists then $\int_{-\infty}^{\infty} f_{\eta_1 \eta_2}(x, y) dx = f_{\eta_2}(y)$.

DEFINITION. The covariance of two random variables η_1 and η_2 is

$$\mathrm{Cov}(\eta_1, \eta_2) = E[(\eta_1 - E[\eta_1])(\eta_2 - E[\eta_2])].$$

If $\mathrm{Cov}(\eta_1, \eta_2) = 0$ then the random variables are uncorrelated. It is in general not true that uncorrelated variables are independent.

EXAMPLE. Let η_1 and η_2 be two random variables with joint probability distribution

$$(\eta_1, \eta_2) = \begin{cases} (\frac{1}{2}, \frac{1}{4}), & \text{with probability } \frac{1}{4} \\ (\frac{1}{2}, -\frac{1}{4}), & \text{with probability } \frac{1}{4} \\ (-\frac{1}{2}, 0), & \text{with probability } \frac{1}{2} \end{cases}.$$

Then we have $E[\eta_1] = 0$, $E[\eta_2] = 0$, and $E[\eta_1 \eta_2] = 0$. However, the random variables are not independent.

Finally, a vector-valued random variable is Gaussian (or equivalently a sequence of random variables is jointly Gaussian) if

$$P(x_1 \leq \eta_1 \leq x_1 + dx_1, \ldots, x_n \leq \eta_n \leq x_n + dx_n) = \frac{1}{Z} e^{-\frac{1}{2}((x-m)A^{-1}(x-m))}$$

where $x = (x_1, x_2, \ldots, x_n)$, $m = (m_1, m_2, \ldots, m_n)$ and A is a positive definite $m \times m$ matrix. The normalization constant Z can be shown to be $Z = (2\pi)^{n/2}|A|^{1/2}$ where $|A|$ is the determinant of A. For the case of jointly Gaussian random variables the covariance matrix C with entries $C_{ij} = E[(\eta_i - E[\eta_i])(\eta_j - E[\eta_j])]$ is the matrix A. If $C_{ij} = 0$ then η_i and η_j are uncorrelated. Furthermore, two Gaussian variables that are uncorrelated are also independent.

3. Chebyshev's Inequality and the Weak Law of Large Numbers

We now discuss several useful properties of the mathematical expectation E.

LEMMA 2.1. $E[\eta_1 + \eta_2] = E[\eta_1] + E[\eta_2]$.

PROOF. We assume for simplicity that the joint density $f_{\eta_1\eta_2}(x, y)$ exists. Then the density $f_{\eta_1}(x)$ of η_1 is given by

$$f_{\eta_1}(x) = \int_{-\infty}^{\infty} f_{\eta_1\eta_2}(x, y)dy$$

and the density $f_{\eta_2}(y)$ of η_2 is given by

$$f_{\eta_2}(y) = \int_{-\infty}^{\infty} f_{\eta_1\eta_2}(x, y)dx,$$

therefore

$$\begin{aligned} E[\eta_1 + \eta_2] &= \int (x + y) f_{\eta_1\eta_2}(x, y)dxdy \\ &= \int x f_{\eta_1\eta_2}(x, y)dxdy + \int y f_{\eta_1\eta_2}(x, y)dxdy \\ &= \int xdx \int f_{\eta_1\eta_2}(x, y)dy + \int ydy \int f_{\eta_1\eta_2}(x, y)dx \\ &= \int x f_{\eta_1}(x)dx + \int y f_{\eta_2}(y)dy = E[\eta_1] + E[\eta_2]. \end{aligned}$$

■

LEMMA 2.2. *If η_1 and η_2 are independent random variables then*

$$\mathrm{Var}[\eta_1 + \eta_2] = \mathrm{Var}[\eta_1] + \mathrm{Var}[\eta_2].$$

PROOF. For simplicity we assume that η_1 and η_2 have densities with mean zero. Then

$$\begin{aligned}\mathrm{Var}[\eta_1+\eta_2] &= E[(\eta_1+\eta_2-E[\eta_1+\eta_2])^2] = E[(\eta_1+\eta_2)^2] \\ &= \int (x+y)^2 f_{\eta_1\eta_2}(x,y)dxdy \\ &= \int x^2 f_{\eta_1\eta_2}(x,y)dxdy + \int y^2 f_{\eta_1\eta_2}(x,y)dxdy \\ &\quad +2\int xy f_{\eta_1\eta_2}(x,y)dxdy.\end{aligned}$$

The first two integrals above are equal to $\mathrm{Var}(\eta_1)$ and $\mathrm{Var}(\eta_2)$ respectively. The third integral is zero. Indeed, because η_1 and η_2 are independent $f_{\eta_1\eta_2}(x,y) = f_{\eta_1}(x)f_{\eta_2}(y)$ and

$$\int xy f_{\eta_1\eta_2}(x,y)dxdy = \int x f_{\eta_1}(x)dx \int y f_{\eta_2}(y)dy = E[\eta_1]E[\eta_2] = 0.$$

■

Another simple property of the variance is that $\mathrm{Var}(a\eta) = a^2\mathrm{Var}(\eta)$. Indeed,

$$\begin{aligned}\mathrm{Var}(a\eta) &= \int (ax - E[a\eta])^2 f_\eta(x)dx \\ &= \int (ax - aE[\eta])^2 f_\eta(x)dx \\ &= a^2\int (x - E[\eta])^2 f_\eta(x)dx \\ &= a^2\mathrm{Var}(\eta).\end{aligned}$$

We now prove a very useful estimate due to Chebyshev.

LEMMA 2.3. *Let η be a random variable. Suppose $g(x)$ is a nonnegative, nondecreasing function, i.e., $g(x) \geq 0$ and $a < b \Rightarrow g(a) \leq g(b)$. Then for any a*

$$P(\eta \geq a) \leq \frac{E[g(\eta)]}{g(a)}.$$

PROOF.

$$\begin{aligned}E[g(\eta)] = \int_{-\infty}^{\infty} g(x)f(x)dx &\geq \int_a^{\infty} g(x)f(x)dx \\ &\geq g(a)\int_a^{\infty} f(x)dx = g(a)P(\eta \geq a).\end{aligned}$$

■

EXAMPLE. Suppose η is a nonnegative random variable. We define $g(x)$ to be 0 when $x \leq 0$ and x^2 when $x \geq 0$. Let a be any positive number. Then

$$P(\eta \geq a) \leq \frac{g(\eta)}{a^2} = \frac{E[\eta^2]}{a^2}.$$

Consider now a special case. Let η be a random variable and define $\xi = |\eta - E[\eta]|$. Then we obtain the following inequality

$$P(|\eta - E[\eta]| \geq a) \leq \frac{\mathrm{Var}(\eta)}{a^2}$$

for any $a > 0$. Now take $a = \sigma k$ where k is an integer. Then

$$P(|\eta - E[\eta]| \geq \sigma k) \leq \frac{\mathrm{Var}(\eta)}{(\sigma k)^2} = \frac{1}{k^2}.$$

In other words, it is very unlikely that η differs from its expected value by more than a few standard deviations.

Suppose $\eta_1, \eta_2, \ldots, \eta_n$ are independent, identically distributed random variables. Let

$$\eta = \frac{1}{n} \sum_{i=1}^{n} \eta_i.$$

Then

$$E[\eta] = E[\eta_1], \quad \mathrm{Var}(\eta) = \frac{1}{n}\mathrm{Var}(\eta_1), \quad \sigma(\eta) = \frac{\sigma(\eta_1)}{\sqrt{n}}.$$

Therefore

$$P\left(|\eta - E[\eta]| \geq k n^{-1/2} \sigma(\eta_1)\right) \leq \frac{1}{k^2}.$$

This tells us that if we use the average of n independent samples of a given distribution to estimate the mean of the distribution then the error in our measurement decreases as $1/\sqrt{n}$. This brings the notion of expected value closer to the intuitive, every day notion of "average."

4. Monte Carlo Methods

With Monte Carlo methods one evaulates a non-random quantity as an expected value of a random variable.

A pseudo-random sequence is a computer generated sequence with independent elements which cannot be distinguished by simple tests from a random sequence, yet is the same each time one runs the appropriate program. For the equidistribution density, number theory allows us to construct the appropriate pseudo-random sequence. Suppose that we want to generate a sequence of independent pseudo-random numbers with probability distribution function $F(x)$. This can be done in the following way. Let $F(\eta) = \xi$ where η is the random variable we

want to sample and ξ is equidistributed in $[0,1]$. Take η such that $\eta = F^{-1}(\xi)$ holds (if there are multiple solutions pick one arbitrarily). Then η will have the desired distribution. To see this consider the following example. Let η be a random variable with

$$\eta = \begin{cases} \alpha_1 \text{ with probability } p_1 \\ \alpha_2 \text{ with probability } p_2 \\ \alpha_3 \text{ with probability } p_3 \end{cases},$$

where $\sum_{i=1}^{3} p_i = 1$ and $p_i \geq 0$ for $i = 1, 2, 3$. Then $F(\eta) = \xi$ implies

$$\eta = \begin{cases} \alpha_1 \text{ if } \xi \in [0, p_1], \\ \alpha_2 \text{ if } \xi \in (p_1, p_2], \\ \alpha_1 \text{ if } \xi \in (p_2, 1]. \end{cases}$$

This can be generalized to any countable number of discrete values in the range of η, and since any function can be approximated by a step function, the results hold for any probability distribution function F.

Example. Let η be a random variable with the exponential pdf. Then $F(\eta) = \xi$ gives

$$\int_0^\eta e^{-s} ds = \xi \implies \eta = -\log(1 - \xi).$$

Example. If f exists then by differentiating $\int_{-\infty}^{\eta} f(s)ds = \xi$ we get $f(\eta)d\eta = d\xi$. The following algorithm (Box-Muller) allows us to sample pairs of independent variables with Gaussian densities with zero mean and variance σ^2. If

$$\eta_1 = \sqrt{-2\sigma^2 \log \xi_1} \cos(2\pi\xi_2)$$
$$\eta_2 = \sqrt{-2\sigma^2 \log \xi_1} \sin(2\pi\xi_2)$$

where ξ_1 and ξ_2 are equidistributed in $[0,1]$, as one can see from

$$\begin{vmatrix} \frac{\partial \eta_1}{\partial \xi_1} & \frac{\partial \eta_1}{\partial \xi_2} \\ \frac{\partial \eta_2}{\partial \xi_1} & \frac{\partial \eta_2}{\partial \xi_2} \end{vmatrix}^{-1} d\eta_1 d\eta_2 = d\xi_1 d\xi_2$$

thus

$$\frac{\xi_1}{2\pi} d\eta_1 d\eta_2 = \frac{1}{2\pi} \exp\left(-\frac{\eta_1^2 + \eta_2^2}{2\sigma^2}\right) d\eta_1 d\eta_2 = d\xi_1 d\xi_2.$$

Now we present the Monte Carlo method. Consider the problem of evaluating the integral $I = \int_a^b g(x)f(x)dx$, where $f(x) \geq 0$ and

$\int_a^b f(x)dx = 1$. We have

$$I = \int_a^b g(x)f(x)dx = E[g(\eta)]$$

where η is a random variable with pdf $f(x)$. Suppose that we can sample η, i.e., make n independent experiments and find values $\eta_1, \ldots, \eta_n$. Then, as can be seen from the Chebyshev inequality, we can approximate $E[g(\eta)]$ by

$$E[g(\eta)] = \frac{1}{n}\sum_{i=1}^{n} g(\eta_i).$$

The error in this approximation will be of the order of $\sigma(g(\eta))/\sqrt{n}$, where $\sigma(g(\eta))$ is the standard deviation of the variable $g(\eta)$. The integral I is the estimand, $g(\eta)$ is the estimator, and $n^{-1}\sum_{i=1}^{n} g(\eta_i)$ is the estimate. The estimator is unbiased if its expected value is the estimand.

Example. Let

$$I = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x)e^{-x^2/2}dx.$$

If η is a Gaussian random variable with mean 0 and variance 1 then

$$I = E[g(\eta)] \sim \frac{1}{n}\sum_{i=1}^{n} g(\eta_i).$$

There are two ways to reduce the error of a Monte Carlo method as can be seen from the error estimate. One way is to take a larger number of samples. The other way is to reduce the variance of the function $g(\eta)$. One way to reduce the variance is "importance sampling."

We start with an extreme case. Suppose we want to evaluate the integral $I = \int_a^b g(x)f(x)dx$ as above. Suppose that the function g is nonnegative, then the quantity $q(x)$ given by $q(x) = f(x)g(x)/I$ has the following properties:

$$q(x) \geq 0, \quad \int_a^b q(x)dx = 1.$$

Further, suppose we can generate a pseudo-random sequence with pdf $q(x)$. Then we have

$$\int_a^b g(x)f(x)dx = I\int_a^b \frac{g(x)f(x)}{I}dx = I\int_a^b q(x)dx = IE[1],$$

where 1 is the function that takes the value 1 for all samples. Then the Monte Carlo method has zero error. The problem lies in the definition

of $q(x)$ where we need to know the value of I which is exactly what we want to compute. This shows us that if we know the value of the quantity we want to compute then Monte Carlo can give us the exact result.

However, it is possible to reduce the error of the Monte Carlo method along similar lines without knowing the result we want to compute. Suppose that we can find a function $h(x)$ with the following properties:

(1) The integral $I_1 = \int_a^b f(x)h(x)dx$ is easily evaluated,
(2) $h(x) \geq 0$,
(3) We can sample a variable with pdf $f(x)h(x)/I_1$ easily,
(4) $g(x)/h(x)$ varies little.

Then we have

$$I = \int_a^b g(x)f(x)dx = \int_a^b \frac{g(x)}{h(x)} f(x)h(x)dx = I_1 \int_a^b \frac{g(x)}{h(x)} \frac{f(x)h(x)}{I_1} dx$$
$$= I_1 E\left[\frac{g}{h}(\eta)\right] \sim \frac{I_1}{n} \sum_{i=1}^{n} \frac{g(\eta_i)}{h(\eta_i)} \quad (2.2)$$

where η has pdf $f(x)g(x)/I_1$. Since $g(\eta)/h(\eta)$ varies little, its variance and the error will be smaller. For a more quantitative estimate see later. Note that the new random variable puts more points where g is large, hence the name of the method "importance sampling"; one puts more samples where g is large, or "important."

Example. Suppose that we want to compute via Monte Carlo the integral $I = \int_0^1 \cos(x/5)e^{-5x}dx$. We can do that by application of the basic Monte Carlo formula without any attempt at importance sampling. That would mean to sample (independently) a variable ξ n times and then approximate I by

$$I \approx \frac{1}{n} \sum_{i=1}^{n} \cos(\xi_i/5)e^{-5\xi_i}.$$

However, due to the large variance of the function $\cos(x/5)e^{-5x}$, the corresponding error would be large (the large variance of the function is due to the presence of the factor e^{-5x}). Alternatively, we can perform the Monte Carlo integration using importance sampling. There are different ways of doing that and one of them is as follows. Let $I_1 = \int_0^1 e^{-5x}dx$. Then we have

$$I = \int_0^1 \cos(x/5)e^{-5x}dx = I_1 \int_0^1 \cos(x/5)\frac{e^{-5x}}{I_1}dx.$$

Let η be a random variable with pdf

$$f(x) = \begin{cases} \dfrac{e^{-5x}}{I_1}, & 0 \le x \le 1 \\ 0, & \text{elsewhere} \end{cases},$$

then I can be written as $I = I_1 E[\cos(\eta/5)]$. As can be readily seen the function $\cos(x/5)$ has smaller variance in the range of integration $[0, 1]$ than the previous integrand. In order to perform the Monte Carlo integration we need to sample the variable η. As shown above this can be done by solving the equation $\int_0^\eta e^{-5x}/I_1 dx = \xi$, where ξ is equidistributed in $[0, 1]$. An easy calculation gives $\eta = -\frac{1}{5}\log(1-5I_1\xi)$. We can use this formula to sample η n times and thus the Monte Carlo approximation to I will read

$$I \approx \frac{I_1}{n} \sum_{i=1}^{n} \cos(\eta_i/5).$$

5. Parametric Estimation

Suppose η is a random variable which someone has sampled and given you the sample $(x_1, x_2, \ldots, x_n)$. Now try to guess the pdf of the x_i which gave you the sample. Suppose you know the type of distribution you have, but not the parameters of the distribution. For example, suppose you know that the distribution is Gaussian, but you don't know the mean and the variance.

Definition. Any function of a sample is called a "statistic."

Suppose you want to estimate a parameter θ of the pdf by a statistic $\hat{\theta}(x_1, x_2, \ldots, x_n)$.

Definition. The estimate is unbiased if

$$E[\hat{\theta}(x_1, x_2, \ldots, x_n)] = \theta.$$

For example, the sample mean defined by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is unbiased, while the sample variance

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is not unbiased. But one can check that

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

is an unbiased estimate of the variance.

Suppose you think that the pdf of η which gave you the independent sample $x=(x_1,x_2,\dots,x_n)$ is $f(\theta)$. Then the following question arises: what is a good estimate of θ given x? Suppose you know θ. Then the probability of getting the given sample is

$$L=\prod_{i=1}^{n} f(x_i \mid \theta).$$

L is called a likelihood function. It is plausible that a good estimate of θ is the one that maximizes L. This is the "maximum likelihood estimate." In general it is easier to maximize $\log L$.

EXAMPLE. Suppose you think that $x_1,x_2,\dots,x_n$ are independent samples of a Gaussian distribution with mean m and variance σ^2. Then

$$L=\prod_{i=1}^{n}\frac{e^{-(x_i-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

Find the maximum of $\log L$:

$$\log L=\sum_{i=1}^{n}\left(-\frac{(x_i-m)^2}{2\sigma^2}-\frac{1}{2}\log 2\pi-\log\sigma\right),$$

$$\frac{\partial \log L}{\partial m}=\sum_{i=1}^{n}\frac{x_i-m}{\sigma^2}=0.$$

Hence

$$\sum_{i=1}^{n}x_i-nm=0,$$

and we get the sample mean as the maximum likelihood estimate of m:

$$m=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

Similarly,

$$\frac{\partial \log L}{\partial \sigma}=-\frac{n}{\sigma}+\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^3}=0,$$

hence the maximum likelihood estimate of the variance of a Gaussian variable is the sample variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2.$$

6. The Central Limit Theorem

Suppose that $\eta_1, \eta_2, \ldots, \eta_n$ are independent, identically distributed random variables. We can assume without loss of generality that they have mean zero and variance one. Suppose the η_i's have a pdf f. Define a new random variable

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i.$$

What can we say about the pdf of S_n? The answer for this question is given by

THEOREM 2.4. *(**The Central Limit Theorem**) Let $\eta_1, \eta_2, \ldots, \eta_n$ be independent, identically distributed random variables with finite variance and zero mean. Let us also assume for simplicity that* $\mathrm{Var}(\eta_i) = 1$. *Then*

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i$$

converges weakly to a Gaussian variable with mean zero and variance one.

PROOF. We will assume that the η_i have pdf f and that f_s is the pdf of S_n. We want to show that

$$\lim_{n\to\infty}\int_a^b f_s(x)dx = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx$$

for any a, b. Note that $n^{-1}\sum\eta_i = n^{-1/2}(n^{-1/2}\sum\eta_i)$ where $n^{-1/2}\sum\eta_i$ tends to a Gaussian; thus the central limit theorem contains information as to how $n^{-1}\sum\eta_i \to 0$ (i.e., for large n, $n^{-1}\sum\eta_i \approx \text{Gaussian}/\sqrt{n}$). Suppose η_1 and η_2 are random variables with respective pdf's f_1 and f_2. What is the density of $\eta_1 + \eta_2$? We know that

$$P(\eta_1 + \eta_2 \le x) = F_{\eta_1+\eta_2}(x) = \int\int_{x_1+x_2\le x} f_1(x_1)f_2(x_2)dx_1dx_2.$$

With the change of variables $x_2 = t$ and $x_1 + x_2 = x$ (note that the Jacobian is 1), we obtain:

$$F_{\eta_1+\eta_2}(x) = \int dx \int f_1(t)f_2(x-t)dt.$$

Thus the density of $\eta_1+\eta_2 = f_{\eta_1+\eta_2}$ is just $\int f_1(x_2)f_2(x-x_2)dx_2 = f_1 * f_2$ and hence $\hat{f}_{\eta_1+\eta_2} = \hat{f}_1\hat{f}_2$.

Hence if we assume that the random variables η_i have the same density function for all i, then $\sum_{i=1}^n \eta_i$ has density $f^{(n)} = f * f * \ldots * f$ n times, where $*$ is the convolution. Furthermore,

$$P(a < S_n \leq b) = P\left(a < \frac{1}{\sqrt{n}}\sum \eta_i \leq b\right) = P(\sqrt{n}a \leq \sum \eta_i \leq \sqrt{n}b)$$
$$= \int_{\sqrt{n}a}^{\sqrt{n}b} f^{(n)}(x)dx = \int_a^b \sqrt{n}f^{(n)}(y\sqrt{n})dy. \quad (2.3)$$

The last step involves the change of variables $y = x/\sqrt{n}$.

What we want to show is that $\int_a^b \sqrt{n}f^{(n)}(y\sqrt{n})dy$ converges to

$$\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx.$$

Pick some nice function ϕ and consider

$$I = \int_{-\infty}^{\infty} f_s(x)\phi(x)dx = \int_{-\infty}^{\infty} \sqrt{n}f^{(n)}(x\sqrt{n})\phi(x)dx.$$

Let $\check{\phi}(k)$ be the inverse Fourier transform of ϕ, i.e.,

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \check{\phi}(k)e^{-ikx}dk.$$

Then

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \sqrt{n}f^{(n)}(x\sqrt{n})\phi(x)dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sqrt{n}f^{(n)}(x\sqrt{n})dx \int_{-\infty}^{\infty} \check{\phi}(k)e^{-ikx}dk \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \sqrt{n}f^{(n)}(x\sqrt{n})e^{-ikx}dx\right)\check{\phi}(k)dk \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \left[\hat{f}\left(\frac{k}{\sqrt{n}}\right)\right]^n \check{\phi}(k)dk. \end{aligned}$$

Here

$$\hat{f}\left(\frac{k}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{ikx/\sqrt{n}}dk.$$

Expand $e^{ikx/\sqrt{n}}$ in a Taylor series:

$$e^{ikx/\sqrt{n}} = 1 + \frac{ixk}{\sqrt{n}} - \frac{x^2k^2}{2n} + O\left(\frac{1}{n^{3/2}}\right).$$

Then

$$f(x)e^{ikx/\sqrt{n}} = f(x) + \frac{ixk}{\sqrt{n}}f(x) - \frac{x^2k^2}{2n}f(x) + O\left(\frac{1}{n^{3/2}}\right).$$

Recall that

$$\int f(x)dx = 1, \quad \int xf(x)dx = 0, \quad \int x^2 f(x)dx = 1.$$

Hence

$$\hat{f}\left(\frac{k}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(1 - \frac{k^2x^2}{2n} + \ldots\right)f(x)dx = 1 - \frac{k^2}{2n} + \text{small}.$$

Since

$$\lim_{n\to\infty}\left(1 - \frac{a}{n}\right)^n = e^{-a}$$

and since the integral of the small terms of the expansion is negligible we get

$$\lim_{n\to\infty}\left[\hat{f}\left(\frac{k}{\sqrt{n}}\right)\right]^n = \lim_{n\to\infty}\left(1 - \frac{k^2}{2n} + \text{small}\right)^n = e^{-k^2/2}.$$

Returning to the integral I we obtain

$$\begin{aligned} I &\to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-k^2/2}\,\check{\phi}(k)dk \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-k^2/2}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(x)e^{-ikx}dx\right)dk \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(x)dx\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-k^2/2}e^{-ikx}dk\right) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(x)e^{-x^2/2}dx. \end{aligned}$$

Now, taking ϕ to be a smooth function that approximates

$$\phi(x) = \begin{cases} 1, & a \le x \le b \\ 0, & \text{otherwise} \end{cases},$$

we get the desired result. ■

7. Conditional Probability and Conditional Expectation.

Suppose we make an experiment and observe that event A has happened, with $P(A) \neq 0$. How does this knowledge affect the probability that event B happens also? We define the probability of B given A to be

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

If A and B are independent then $P(A \cap B) = P(A)P(B)$ and so

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B).$$

If A is fixed and B is any member of $\mathcal{B}$ (i.e., any event) than $P(B|A)$ defines a perfectly good probability on $\mathcal{B}$; this is the probability conditional on A

$$(\Omega, \mathcal{B}, P) \to (\Omega, \mathcal{B}, P(B|A)).$$

Suppose η is a random variable on Ω. Then the average of η given A is

$$E[\eta|A] = \int \eta(\omega) P(d\omega|A).$$

EXAMPLE. Suppose we throw a die. Let η be the value on top. Then

$$E[\eta] = \frac{1}{6} \sum_{i=1}^{6} i = 3.5.$$

Suppose we know that the outcome is odd. Then the probability that the outcome is one is given by

$$P(1) = \frac{P(\{1\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

In this case the conditional expectation of η given A is

$$E[\eta|\text{outcome is odd}] = \frac{1}{3}(1+3+5) = 3.$$

The probabilities of an even outcome is

$$P(2) = P(4) = P(6) = 0$$

while the total probability of an odd outcome is

$$P(1) + P(3) + P(5) = 1.$$

Suppose $Z = \{Z_i\}$ is an at most countable disjoint measurable partition of Ω. This means that the number of Z_i's is finite or countable, each Z_i is an element of $\mathcal{B}$, $\Omega = \bigcup_i Z_i$, and $Z_i \cap Z_j = \emptyset$ if $i \neq j$.

EXAMPLE. $Z = \{A, CA\}$, where A is a measurable subset of Ω and CA is the complement of A.

DEFINITION. Suppose A is an event. Then $\mathcal{X}_A(\omega)$ is a random variable equal to 1 when $\omega \in A$ and 0 when $\omega \notin A$.

Observe that $E[\mathcal{X}_A(\omega)] = P(A)$.

DEFINITION. Let $Z = \{Z_i\}$ be a partition of Ω as above. Let η be a random variable and construct the random variable $E[\eta|Z]$ as follows

$$E[\eta|Z] = \sum_i \mathcal{X}_{Z_i} E[\eta|Z_i].$$

This is a function of ω whose definition depends on the choice of partition Z. In words, we average η over each element Z_i of the partition and then we assign this average to be the value of the variable $E[\eta|Z]$ for all ω in Z_i. If one could think of the elements of ω as people and the values of η as those people's heights, one could then partition the people by ethnic origin and assign an average height to each ethnic group. Given a person, the new variable would consider that person's height to be the average height of his ethnic group.

Note that Z generates a σ-algebra. It is a coarser σ-algebra than $\mathcal{B}$, i.e., it is contained in $\mathcal{B}$. The conditional expectation is the best estimate of the original random variable when the instruments you use to measure the outcomes (which define the σ-algebra $\mathcal{B}$) are too coarse.

EXAMPLE. Return to the example of the die. Let η be the number on top. Let A be the event that outcome is odd. Let $Z = \{A, CA\}$. Then

$$E[\eta|A] = \frac{1}{3}(1+3+5) = 3,$$

$$E[\eta|CA] = \frac{1}{3}(2+4+6) = 4,$$

and finally

$$E[\eta|Z] = 3\mathcal{X}_A + 4\mathcal{X}_{CA}.$$

We now want to define the notion of conditional expectation of one random variable η given another random variable ξ. For simplicity we assume at first that ξ takes only finitely many values $\xi_1, \xi_2, \ldots, \xi_n$. Let Z_i be the inverse image of ξ_i (the set of ω such that $\eta(\omega) = \xi_i$). Then $Z = \{Z_1, Z_2, \ldots, Z_n\}$ is a finite disjoint partition of Ω. Thus we can construct $E[\eta|Z]$ as defined above.

DEFINITION. We define $E[\eta|\xi]$ to be equal to the random variable $E[\eta|Z]$.

We observe that $E[\eta|\xi]$ is a random variable and at the same time a function of ξ. Indeed, when ξ has value ξ_i then $E[\eta|\xi] = E[\eta|Z_i]$, thus $E[\eta|\xi]$ is a function of ξ. We now show that $E[\eta|\xi]$ is actually the best least square approximation of η by a function of ξ. This property can serve as an alternative definition of conditional expectation.

THEOREM 2.5. *Let $h(\xi)$ be any function of ξ. Then*

$$E[(\eta - E[\eta|\xi])^2] \leq E[(\eta - h(\xi))^2].$$

PROOF. We remind the reader that

$$\int_0^1 (f(x) - c)^2 dx$$

is minimized when c is the average of $f(x)$ on $[0,1]$, that is, when $c = \int_0^1 f(x)dx$. Similarly, we want to minimize

$$\begin{aligned} E[(\eta - g(\xi))^2] &= \int_\Omega (\eta - g(\xi(\omega))^2 P(d\omega) \\ &= \sum_i P(Z_i) \int_{Z_i} (\eta - g(\xi(\omega))^2 \frac{P(d\omega)}{P(Z_i)}. \end{aligned}$$

Each of the integrals $\int_{Z_i}(\eta - g(\xi(\omega))^2 P(d\omega)/P(Z_i)$ is minimized when $g(\xi(\omega)) = E[\eta|Z_i]$, that is when $g(\xi(\omega))$ is the average of η on Z_i. Thus $E[\eta|\xi]$ is the best least squares approximation of η by a function of ξ. ■

Consider the space of all random variables. It is a vector space. The random variables which are functions of ξ form a linear subspace. Let η_1, and η_2 be random variables. Define their scalar product by

$$(\eta_1, \eta_2) = E[\eta_1 \eta_2].$$

The space of functions of ξ is closed in the norm

$$||\eta|| = \sqrt{(\eta, \eta)}.$$

Theorem 2.5 implies that $E[\eta|\xi]$ is the orthogonal projection of η on the space of functions of ξ. This in turn implies that for any function $g(\xi)$ we have

$$E[(\eta - E[\eta|\xi])g(\xi)] = 0.$$

The meaning of the formula above is that $\eta - E[\eta|\xi]$ is perpendicular to all functions of ξ in the space of functions which are square integrable with respect to the probability P, i.e., have finite variance. If we define $\mathbb{P}$ to be the orthogonal projection which projects the whole space of random variables with finite variance onto the subspace of random variables with finite variance which are functions of ξ, then $E[\eta|\xi] = \mathbb{P}\eta$.

We now consider the special case where η and ξ are random variables whose joint density $f_{\eta\xi}$ is known

$$P(s < \eta \leq s + ds, t < \xi \leq t + dt)) = f_{\eta\xi}(s,t)dsdt.$$

We want to calculate $E[g(\eta, \xi)|\xi]$ where $g(\eta, \xi)$ is some function of η and ξ. $E[g(\eta, \xi)|\xi]$ is a random variable and a function of ξ. What is

this function? Specifically, what is the value of this random variable when $\xi = a$?

To answer this question, we first define a discrete approximation $\hat{\xi}$ to ξ which takes the value $\hat{\xi} = (i + 1/2)h$ when $\xi \in [ih, (i + 1)h]$. This happens with probability $\int_{ih}^{(i+1)h} f_\xi(t)dt$ where $f_\xi(t)$ is given by

$$f_\xi(t) = \int_{-\infty}^{\infty} f_{\eta\xi}(s, t)ds.$$

Now we replace $E[g(\eta, \xi)|\xi]$ by $E[g(\eta, \xi)|\hat{\xi}]$. (We are commiting many mathematical sins here, but sin should be enjoyed.) Suppose we fix an a and pick a value $a_i = (i + 1/2)h$ of $\hat{\xi}$ such that $a \in [ih, (i + 1)h]$. Then

$$\begin{aligned} E[g(\eta, \xi)|\xi]_{\xi=a} &\approx E\left[g(\eta, \xi)|\hat{\xi}\right]_{\hat{\xi}=a_i} \\ &\approx \frac{\int_{-\infty}^{\infty} ds \int_{ih}^{(i+1)h} g(s, t)f(s, t)dt}{\int_{-\infty}^{\infty} hf(s, (i + 1/2))ds} \to \frac{\int_{-\infty}^{\infty} g(s, a)f(s, a)ds}{\int_{-\infty}^{\infty} f(s, a)ds} \end{aligned}$$

as $h \to 0$. Thus

$$E[g(\eta, \xi)|\xi]_{\xi=a} = \frac{\int_{-\infty}^{\infty} g(s, a)f(s, a)ds}{\int_{-\infty}^{\infty} f(s, a)ds}.$$

This is just what one would expect: the value of $E[g(\eta, \xi)|\xi]$ when $\xi = a$ is the mean of $g(\eta, \xi)$ when we keep ξ equal to a but allow η to take any value it wants.

8. Conditional Probabilities and Bayes' Theorem

Recall the definition of conditional probability:

DEFINITION. Let A and B be two events with $P(A) \neq 0$ and $P(B) \neq 0$. The conditional probability of B given A, $P(B|A)$, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.4}$$

Similarly, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.5}$$

Combining (2.4) and (2.5) we get Bayes' theorem:

THEOREM 2.6. *Let A and B be two events with $P(A) \neq 0$ and $P(B) \neq 0$. Then*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.6}$$

Suppose $Z = \{Z_j\}, j = 1, 2, \ldots$ is a finite or countable partition of the sample space Ω as above; then for the probability $P(A)$ of an event A we have

$$P(A) = \sum_j P(A \cap Z_j) = \sum_j \frac{P(A \cap Z_j)}{P(A)} P(A) = \sum_j P(Z_j|A)P(A).$$

Suppose that $P(Z_j) \neq 0$ for all j. Then we can also rewrite $P(A)$ as

$$P(A) = \sum_j P(A \cap Z_j) = \sum_j \frac{P(A \cap Z_j)}{P(Z_j)} P(Z_j) = \sum_j P(A|Z_j)P(Z_j). \tag{2.7}$$

Using Bayes' theorem (2.6) for the events A and Z_j and expressing $P(A)$ by (2.7) we get

$$P(Z_j|A) = \frac{P(A|Z_j)P(Z_j)}{\sum_j P(A|Z_j)P(Z_j)}. \tag{2.8}$$

This is the second form of Bayes' theorem. We can use the second form to address the following question: Suppose we have an experimental sample and we know that we have sampled some probability distribution which depends on a parameter θ. We do not know what value θ takes in the case at hand, but we have an idea à priori (i.e., a "prior" idea) that the set of possible values of θ can be viewed as a random variable with a density g_{old} (the "prior" distribution). Now that we have made an experiment and obtained data, we should be able to learn from these data how to improve the prior ideas and obtain a new density g_{new}, the "posterior" density, which improves the "prior" density in light of the data. We show how to do it in an example.

EXAMPLE. Let η_1 and η_2 be two independent, identically distributed random variables with

$$\eta_1 = \eta_2 = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}.$$

For the sum $\eta_1 + \eta_2$ we can deduce:

$$\eta_1 + \eta_2 = \begin{cases} 2, & \text{with probability } p^2 \\ 1, & \text{with probability } 2p(1-p) \\ 0, & \text{with probability } (1-p)^2 \end{cases}.$$

Suppose that before the experiment we thought that the parameter p had the value $p = 1/4$ with probability $1/4$ and the value $p = 1/2$ with probability $3/4$. This is the "prior distribution." Now we make an experiment and find $\eta_1 + \eta_2 = 1$. We want to use the second form

of Bayes' theorem (2.8) to see how the experiment affects our beliefs about the distribution of the parameter p. To do that let A be the event that $\eta_1 + \eta_2 = 1$, let Z_1 be the event that $p = 1/4$, and Z_2 the event that $p = 1/2$ (note that $Z_1 \cup Z_2 = \Omega$). Then we have

$$\begin{aligned} P(Z_1|A) &= \frac{P(A|Z_1)P(Z_1)}{\sum_j P(A|Z_j)P(Z_j)} \\ &= \frac{\left(2 \times \frac{1}{4} \times \frac{3}{4}\right) \times \frac{1}{4}}{\left(2 \times \frac{1}{4} \times \frac{3}{4}\right) \times \frac{1}{4} + \left(2 \times \frac{1}{2} \times \frac{1}{2}\right) \times \frac{3}{4}} \\ &= \frac{1}{5}, \end{aligned}$$

as opposed to $1/4$ à priori. In words, the probability that $p = 1/4$ now that we know the outcome of the experiment equals the ratio of the product of the probability that the outcome is what it is when $p = 1/4$ and the prior probability that $p = 1/4$, normalized by the sum of the probabilities of the outcome we have for the various prior probabilities.

Of course the taint of possible error in the prior ideas has not completely disappeared.

9. References

1. T. Amemiya, *Introduction to Statistics and Econometrics*, Harvard Univ. press, Cambridge, (1994).
2. P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*.
3. K.L. Chung, *A Course in Probability Theory, Academic*, NY, (1974).
4. J. Lamperti, *Probability*, Benjamin, NY, 1966.
5. H. Dym and H.P. McKean, *Fourier Series and Integrals*, Academic Press, New York, (1972).